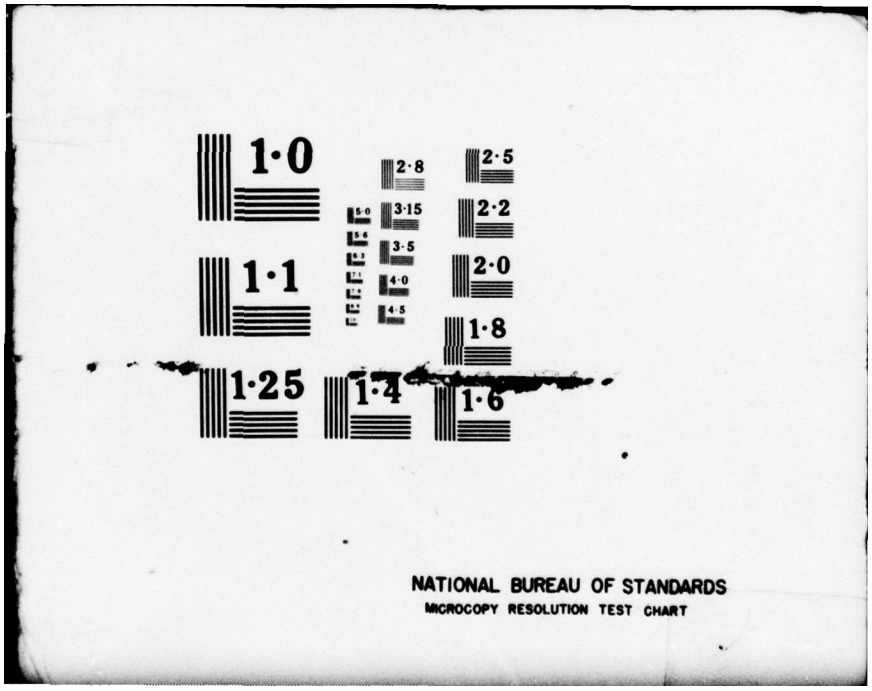
AD-A053 919 TEXAS A AND M UNIV COLLEGE STATION F/G 12/1

AN EXACT SMALL SAMPLE THEORY FOR POST-STRATIFICATION.(U)

MAR 78 D C DOSS, H O HARTLEY, G R SOMAYAJULU DAAG29-77-G-0086

UNCLASSIFIED TR-1 ARO-14209.1-M NL

| OF |
ADA
053919


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

ARO 14209.1-M

12

AD A053919

ARO-D PROJECT DAAG29-77-G-0086

Technical Report No. 1

AN EXACT SMALL SAMPLE THEORY FOR POST-STRATIFICATION

by

D. C. Doss, H. O. Hartley, and G. R. Somayajulu

AD No. ____
DDC FILE COPY

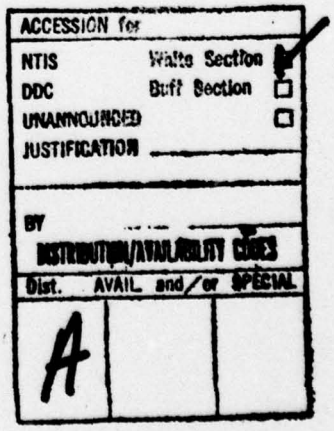
DDC
RECEIVED
MAY 11 1978
B

March 1978

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

ABSTRACT

A genuine small sample theory for post-stratification is developed in this paper. This includes the definition of a ratio estimator of the population mean $\overline{Y}$, the derivation of its bias and its exact variance and a discussion of variance estimation. The estimator has both a within strata component of variance which is comparable with that obtained in proportional allocation stratified sampling and a between strata component of variance which will tend to zero as the overall sample size becomes large. Certain optimality properties of the estimator are obtained. The generalization of post-stratification from the simple random sampling to post-stratification used in conjunction with stratification and multi-stage designs is discussed.



"An Exact Small Sample Theory for Post-Stratification"

by

D. C. Doss*, H. O. Hartley† and G. R. Somayajulu$

1. Introduction

As is well known, strata are defined as nonoverlapping and exhaustive subsets of the units of a population with the following properties:

(a) The total number of units N_h in stratum h of the population is known,

(b) It is possible to identify in advance of sampling the stratum h to which each unit belongs and prescribed sample sizes $n_h \geq 1$ are drawn from stratum h.

"Post strata" differ from strata in the sense that condition (b) is no longer satisfied. However, it is assumed that after sampling it is possible to identify for each elementary unit the post-stratum, h, to which it belongs.

The literature on post-stratification is almost exclusively confined to the case of a simple random sample of size n drawn from the population. If we define by n_h the number of units which "happen to fall" into a post-stratum h then the n_h become random variables following a hypergeometric distribution. It is well known that the literature on post-stratification is essentially confined to a situation where the probability that $n_h = 0$

---

*D. C. Doss, University of Alabama in Huntsville

†H. O. Hartley, Institute of Statistics, Texas A&M University

$G. R. Somayajulu now at Osmania University, Hyderabad, INDIA. The initial stages of this work formed part of the dissertation by G. R. S. prepared at Texas A&M University under the direction of Dr. H. O. Hartley.

can be assumed to be negligible. Accordingly, the estimator of the population mean $\bar{Y}$ considered is of the form

$$\hat{y} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}'_h \tag{1.1}$$

where the N_h/N are the known post-strata proportions in the population and $\bar{y}'_h$ is the ordinary sample mean of the units falling into post-stratum h provided $n_h \geq 1$. The definition of $\bar{y}'_h$ for the case $n_h = 0$ varies. As is well known, if the probability that $n_h = 0$ is negligibly small the above estimator is approximately unbiased and has a variance which is approximately equal to that of a stratified estimator for proportional allocations.

If the above approximations are accepted it would follow that all the benefits derivable from stratification and proportional allocation can be attained by the above device of post-stratification. Unfortunately, experience with post-stratification when the sample size n is comparatively small and the number of strata is comparatively large is distinctly disappointing.

It is therefore the purpose of this paper to develop a genuine small sample theory for post-stratification. This will include the precise definition of the estimator of $\bar{Y}$, the derivation of its bias and its exact variance and a discussion of variance estimation. It is not surprising that our findings will show that our post-stratified estimator will have both a within strata component of variance which is comparable with that obtained in proportional allocation stratified sampling but also a between strata component of variance which will tend to 0 as the overall sample size n becomes large to an order which is $O(n^{-1})$. The derivation of our

compact and exact variance formulas for both components of variance enables us to derive certain optimality properties of our estimator together with recommendations for sampling strategies.

In the last section we also discuss the generalization of post-stratification for survey designs that are more realistic than a simple random sample. These include post-stratification used in conjunction with stratification and multi-stage designs. However, these generalizations are only discussed in generality and not spelled out in detail.

2. A Ratio Estimator

Throughout this paper, we consider only a simple random sample of size n from a population of size N with L strata ($L \geq 2$). However, generalizations of the design are considered in section 6. Defining the "indicator variables"

$$a_h = \begin{cases} 1 & \text{if at least one unit of the sample of size n is in stratum h} \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

we start with an unbiased estimator of $\bar{Y}$ of the form

$$\tilde{y} = \Sigma a_h P_h \bar{y}_h / E(a_h) \tag{2.2}$$

where $P_h = N_h/N$, $E(a_h) = 1 - \binom{N-N_h}{n} / \binom{N}{n}$ and the summation extends over all strata. When $a_h = 0$, $\bar{y}_h$ can be defined arbitrarily as a constant, say, $\bar{Y}_h$, the population mean of stratum h, since the corresponding term in (2.2) is zero. The unbiasedness of this estimator follows from

$$E(\tilde{y}) = \underset{1}{E}(\underset{2}{E}(\tilde{y})) = \underset{1}{E}(\Sigma a_h P_h \bar{Y}_h / E(a_h)) = \Sigma P_h \bar{Y}_h = \bar{Y} \tag{2.3}$$

where $\underset{2}{E}$ is the conditional expectation given $(n_1,\ldots,n_L)$ and $\underset{1}{E}$ is the expectation over $(n_1,\ldots,n_L)$. Similarly, we define variances $\underset{1}{V}$ and $\underset{2}{V}$ and covariances $\underset{1}{Cov}$ and $\underset{2}{Cov}$. Note a similarity of our estimator to the well-known Horvitz-Thompson estimator of $\bar{Y}$ in a random sample with unequal probabilities of selection.

A serious drawback of $\tilde{y}$ is that its variance depends on the origin of the y values. To demonstrate this we consider a translation of each y to y + c where c is an arbitrary constant and the estimator of $\bar{Y} + c$ becomes

$$(y \tilde{+} c) = \Sigma a_h P_h (\bar{y}_h + c)/E(a_h) = \tilde{y} + c\tilde{x} \tag{2.4}$$

where $$\tilde{x} = \Sigma a_h P_h x_h / E(a_h) = \Sigma a_h P_h / E(a_h) \qquad (2.5)$$

and the variable $x_{hi} = 1$ for all population units. Incidentally, $\tilde{x}$ is an unbiased estimator of $\bar{X} = 1$. Now clearly we have for the variance

$$V(\tilde{y + c}) = V(\tilde{y}) + c^2 V(\tilde{x}) + 2cCov(\tilde{y},\tilde{x}) \qquad (2.6)$$

and it is obvious that $V(\tilde{y + c})$ can be made arbitrarily large by increasing c sufficiently. This is due to the fact that $\tilde{x}$ is not a constant.

In order to eliminate the dependence of the variance on the origin of the y values we turn our attention to a ratio estimator of $\bar{Y}$ which is defined by

$$\tilde{R} = (\tilde{y}/\tilde{x})\bar{X} = \tilde{y}/\tilde{x} = \frac{\Sigma a_h P_h \bar{y}_h / E(a_h)}{\Sigma a_h P_h / E(a_h)} \qquad (2.7)$$

The variance of $\tilde{R}$ is unaffected by translation of y values since from (2.2)

$$(\tilde{y + c})/\tilde{x} = (\tilde{y}/\tilde{x}) + c. \qquad (2.8)$$

Now the ratio estimator (2.8) will in general be slightly biased. However in the particular case where all strata proportions P_h are equal, our ratio estimator is shown to be unbiased in Appendix II. In other cases the bias of $\tilde{R}$ as an estimator of $\bar{Y}$ is of the order of magnitude $O(P_h^2 Q_h^n)$ or $O(P_h Q_h^{n+1})$ where $Q_h = 1 - P_h$ (see Appendix I). Therefore, even for a moderate sample size n the bias is negligible provided the P_h are greater than or equal to cn^{-1}.

If the number of strata is large and all P_h are small, while n is moderate, we show in Appendix I that the bias is of order $O(n^1 L^{-2})$ or $O(n^2 L^{-3})$. Once again the bias is negligible. The bias can be exactly evaluated for a small number of strata by direct computation.

3. The Exact Variance of the Ratio Estimator

There are two components of variance resulting from the well-known relation

$$V(\tilde{R}) = \underset{1}{V}(\underset{2}{E}(\tilde{R})) + \underset{1}{E}(\underset{2}{V}(\tilde{R})) \tag{3.1}$$

where again $\underset{2}{E}$ and $\underset{2}{V}$ are conditional expectations and variances given a set of n_h and $\underset{1}{E}$, $\underset{1}{V}$ are expectations and variances over the n_h. The terms $\underset{1}{V}(\underset{2}{E}(\tilde{R}))$ and $\underset{1}{E}(\underset{2}{V}(\tilde{R}))$ are called the between strata component and the within strata component of variance of $\tilde{R}$ and denoted as $V(\tilde{R})_B$ and $V(\tilde{R})_W$ respectively.

First we derive the between strata component in a compact form which requires recasting $\tilde{R}$ in a simple form as

$$\tilde{R} = \Sigma b_h \bar{y}_h \tag{3.2}$$

where
$$b_h = \frac{a_h P_h / E(a_h)}{\sum_{k=1}^{L} a_k P_k / E(a_k)} \tag{3.3}$$

Since $\Sigma b_h = 1$, we obtain for any fixed h

$$\begin{aligned} \sum_{h' \neq h} \mathrm{Cov}(b_{h'}, b_h) &= E\Big(\sum_{h' \neq h} b_{h'} b_h\Big) - E\Big(\sum_{h' \neq h} b_{h'}\Big) E(b_h) \\ &= E\{(1 - b_h) b_h\} - E(1 - b_h) E(b_h) \qquad (3.4) \\ &= -V(b_h). \end{aligned}$$

Since

$$\underset{2}{E}(\tilde{R}) = \Sigma b_h E(\bar{y}_h) = \Sigma b_h \bar{Y}_h \quad , \tag{3.5}$$

we find that, by virtue of (3.4)

$$V(\tilde{R})_B = \underset{1}{V}(\underset{2}{E}(\tilde{R})) = \sum_{h=1}^{L} \underset{1}{V}(b_h)\bar{Y}_h^2 + \sum_{h'\neq h=1}^{L} \underset{1}{Cov}(b_{h'}, b_h)\bar{Y}_{h'}\bar{Y}_h$$

$$= 1/2\{ \sum_{h=1}^{L} \underset{1}{V}(b_h)\bar{Y}_h^2 + \sum_{h=1}^{L} V(b_{h'})\bar{Y}_{h'}^2 + 2 \sum_{h'\neq h=1}^{L} \underset{1}{Cov}(b_{h'}, b_h)\bar{Y}_{h'}\bar{Y}_h \} \tag{3.6}$$

$$= 1/2 \sum_{h'\neq h=1}^{L} \underset{1}{Cov}(b_{h'}, b_h)(\bar{Y}_h^2 + \bar{Y}_{h'}^2 - 2\bar{Y}_h\bar{Y}_{h'})$$

$$= \sum_{h'\neq h=1}^{L} \frac{\{E(b_{h'})E(b_h) - E(b_{h'}, b_h)\}}{2} (\bar{Y}_{h'} - \bar{Y}_h)^2$$

The within strata component of variance $\tilde{R}$ is given by

$$V(\tilde{R})_W = \underset{1}{E}(\underset{2}{V}(\tilde{R})) = \underset{1}{E}(\Sigma b_h^2 \underset{2}{V}(\bar{y}_h)) = \underset{1}{E}\{\Sigma b_h^2((1/n_h) - (1/N_h))S_h^2\}$$

$$= \Sigma E'(b_h^2/n_h)E(a_h)S_h^2 - \Sigma E(b_h^2)S_h^2/N_h \tag{3.7}$$

where E' stands for the conditional expectation given $n_h \geq 1$ and S_h^2 is the population mean square of stratum h, i.e.

$$S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2/(N_h - 1). \tag{3.8}$$

In the particular case where all P_h are equal, the components of variance of $\tilde{R}$ reduce (see Appendix II) to the very simple forms

$$V(\tilde{R})_B = \{E(1/\nu) - (1/L)\}S_B^2 \tag{3.9}$$

and

$$V(\tilde{R})_W = \{E'(1/\nu^2 n_h)E(a_h) - E(1/\nu^2)(L/N)\}\Sigma S_h^2 \tag{3.10}$$

where ν = the number of strata represented in the sample,

$$S_B^2 = \Sigma(\bar{Y}_h^2 - L\bar{Y}^2)/(L - 1),$$

and $$E(a_h) = 1 - \binom{N((L - 1)/L)}{n} / \binom{N}{n}.$$

4. The Efficiency of $\tilde{R}$

In order to reduce the variance of the estimator of $\bar{Y}$ stratified sampling is employed in practice with different allocation schemes of the sample. In particular when the population strata means differ considerably from each other and the patterns of strata variances S_h^2 differ for different content items, the scheme of proportional allocation is used to eliminate this variability. But, if stratified sampling is not possible because of (b), then it is of interest to find an estimator of $\bar{Y}$ based on post-stratification which would minimize the between strata variation and at the same time would not increase the within strata variation "unduly".

A reasonable class of unbiased estimators of $\bar{Y}$ based on post-stratification that can be computed from the sample is given by

$$\Sigma g(n_h) P_h \bar{y}_h / E(g(n_h)) \qquad (4.1)$$

where $g(n_h)$ is any mathematical function defined for all values of n_h and is such that $g(0) = 0$ and $E(g(n_h)) \neq 0$. This class clearly includes the sample mean $\bar{y}$ by letting $g(n_h) = n_h$. All estimators of (4.1) with the exception of $\bar{y}$ suffer from the same drawback as $\tilde{y}$ defined by (2.1), that is, the increase in variance through translation. Hence the logical step to eliminate this effect is to consider ratio estimators analogous to (2.7).

It is shown in Appendix II that the ratio estimators obtained from (4.1) are unbiased and our estimator $\tilde{R}$ given by (2.6) minimizes the between strata component of variance in this generalized class of estimators (4.1) when all strata sizes are equal. In the case where all strata sizes are not equal, there exists no ratio estimator that minimizes the between strata component of variance if $g(n_h)$ is required not to depend on the population strata means or strata variances.

The between strata component of variance of $\tilde{R}$, $V(\tilde{R})_B$, is always smaller than that of $\bar{y}$ and (as is seen in Appendix I) $V(\tilde{R})_B$ is of an exponential order of magnitude $O(P_h^2 Q_h^n)$ or $O(P_h Q_h^{n+1})$ and approaches zero much more rapidly than $V(\bar{y})_B$ which is of the order $O(n^{-1})$.

When the number of strata L is large, all P_h are small, and n is moderately large, it is seen in Appendix I that an approximate $V(\tilde{R})_B$ is of order $O(nL^{-2})$ or $O(n^2L^{-3})$. This implies that in a situation where the usual estimator (1.1) is at its worst, $\tilde{R}$ has a negligible between strata component of variance.

Turning now our attention to the within strata component of variance we consider an approximation to $V(\tilde{R})_W$ since the exact variance is analytically intractable. In Appendix I we show that to terms of order $O(n^{-1})$ we have that

$$V(\tilde{R})_W = \Sigma P_h \sigma_h^2/n(1 - Q_h^n) + \Sigma Q_h \sigma_h^2/n^2(1 - Q_h^n) \qquad (4.2)$$

which clearly approaches the variance of the estimator used in stratification with proportional allocation for large n, i.e.

$$V(\tilde{R})_W \doteq \Sigma(P_h \sigma_h^2/n). \qquad (4.3)$$

The asymptotic result (4.3) is also correct for large L if all P_h are small and n is moderately large.

The relative efficiency of $\tilde{R}$ as compared with the estimator of $\bar{Y}$ employed in stratified sampling with proportional allocation approaches asymptotically 1 if either n is large or L is large (so that all P_h are small) while n is moderately large.

5. The Estimation of the Variance of $\tilde{R}$

An unbiased estimator of $V(\tilde{R})$ is given by

$$V(\tilde{R}) = \sum_{h' \neq h=1}^{L} a_h a_{h'} \frac{\{E(b_{h'})E(b_h) - E(b_{h'}b_h)\}}{2E(a_h a_{h'})} t_{hh'}$$

$$+ \sum_{h=1}^{L} b_h^2 d_h s_h^2 - \sum_{h=1}^{L} b_h^2 s_h^2 / N_h \qquad (5.1)$$

where $t_{hh'}$ is an unbiased estimator of $(\bar{Y}_h - \bar{Y}_{h'})^2$ given $n_h \neq 0$ and $n_{h'} \neq 0$. We may use the estimator

$$t_{hh'} = a_h a_{h'} \{(\bar{y}_h - \bar{y}_{h'})^2 - (s_h^2/n_h) - (s_{h'}^2/n_{h'})\} \qquad (5.2)$$

when we define $s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)/(n_h - 1)$ (5.3)

and $S_{h'}^2$ by replacing h by h' in (5.3). Finally in (5.1) we define d_h by

$$d_h = \begin{cases} 0 & \text{if } n_h = 0 \\ 1/n_h & \text{if } n_h \geq 1. \end{cases} \qquad (5.4)$$

The computation of $\{(\bar{y}_h - \bar{y}_{h'})^2 - (s_h^2/n_h) - (s_{h'}^2/n_{h'})\}$ in (5.2) is only required if both $n_h \geq 1$ and $n_{h'} \geq 1$ since otherwise $a_h a_{h'} = 0$. However, the definitions of s_h^2 and $s_{h'}^2$ in (5.1) and (5.3) require that both $n_h \geq 2$ and $n_{h'} \geq 2$. In case $n_h = 1$ and/or $n_{h'} = 1$ methods of estimating variances from single units per stratum have to be employed (see e.g. Hartley, Rao, and Kiefer (19)).

If the number of strata is not small, then it becomes very tedious to compute $E(b_h)E(b_{h'}) - E(b_{h'}b_h)$ in which case an approximation is provided in Appendix I equations (A-10) to (A-14).

6. Post-stratification for More General Survey Designs

We confine ourselves here to a brief outline of the main general theory of post-stratification for stratified multi-stage designs. We shall utilize the theory of "Domain Estimation" (see e.g. Hartley (1959)) by identifying post strata with "domains of study".

Denote by y_i the characteristic attached to the i^{th} last stage unit and by $\hat{Y}(y_i)$ the standard unbiased estimator of the population total of the y_i. The estimator $\hat{Y}(y_i)$ is a well defined linear function of the y_i in the sample. Define now the domain variables

$$ {}_hy_i = \begin{cases} y_i & \text{if unit i is in domain h} \\ 0 & \text{if unit i is not in domain h} \end{cases} \tag{6.1} $$

and

$$ {}_hx_i = \begin{cases} 1 & \text{if unit i is in domain h} \\ 0 & \text{if unit 1 is not in domain h} \end{cases} \tag{6.2} $$

and consider the subset of samples for which at least one last stage unit falls into domain h. Denote by

$$ \pi_h = \Pr\{\text{at least one last stage unit in domain h}\}. \tag{6.3} $$

For this subset of samples the estimate $\hat{Y}({}_hx_i)$ of the number of units in domain h will be greater than zero since $\hat{Y}({}_hx_i)$ is a linear function of the ${}_hx_i$ with positive coefficients. Accordingly, we can for this subset of samples compute the ratio estimate of the population domain mean in the form

$$ {}_h\hat{\bar{y}} = \frac{\hat{Y}({}_hy_i)}{\hat{Y}({}_hx_i)} \tag{6.4} $$

which will have a "technical bias" given by

$$ \text{Bias}_h\hat{\bar{y}} = -\text{Cov}({}_h\hat{\bar{y}}, \hat{Y}({}_hx_i))(\pi_h/{}_hM) \tag{6.5} $$

where Cov is a conditional covariance applicable to the above subset of samples and ${}_hM$ is the total number of last stage units in domain h. It is reasonable to assume that $\mathrm{Cov}({}_h\hat{y}, \hat{Y}({}_hx_i))$ will be zero or small since the estimate of the mean value of the y characteristic (${}_h\hat{y}$) is unlikely to be correlated with the estimate of the number of units $\hat{Y}({}_hx_i)$ falling into domain h.

We finally turn to the post-stratified estimates of the population mean and define in analogy to (2.2) the post-stratified estimates

$$\tilde{y} = \Sigma a_h P_h (\hat{y}_h/\pi_h), \quad \tilde{x} = \Sigma a_h P_h (1/\pi_h) \tag{6.6}$$

where
$$a_h = \begin{cases} 1 & \text{if at least one last stage unit is in domain h} \\ 0 & \text{if there is not at least one last stage unit in domain h.} \end{cases} \tag{6.7}$$

Finally we define the double ratio estimator

$$\tilde{R} = \tilde{y}/\tilde{x} \tag{6.8}$$

which is our post-stratified estimator of the population mean.

The main difficulty about using (6.6) and (6.8) is the computation of the π_h defined by (6.3) which would require the knowledge of the domain sizes in each last but one stage unit. However for many survey designs it is possible to compute approximations to the π_h as we shall illustrate below:

Assume that the last stage units are sampled with equal probability and without replacement and use the index j to denote the last but one stage units. Denote by p(s) the probability that a sample s of last but one stage units has been drawn by the specified survey design. Denote

by m_j the specified number of last stage units to be drawn from the j^{th} last but one stage unit if in s. Denote by P_{hj} the proportion of last stage units in the j^{th} last but one stage unit which are in domain h and by $Q_{hj} = 1 - P_{hj}$. Then (ignoring fpc's) the probability π_h is given by

$$\pi_h = 1 - \sum_s p(s) \prod_{j \text{ in } s} Q_{hj}^{mj} \doteq 1 - Q_h^{\bar{m}} \tag{6.9}$$

where Q_h is an average value of the Q_{hj} and $\bar{m}$ is an average value of the total overall last stage sample size.

Improvements in the computation of the π_h and the spelling out of the bias and variance of $\tilde{R}$ will be left to subsequent communications.

APPENDIX I

1. The order of magnitude of the bias and $V(R)_B$ as $n \to \infty$

When the numerator of b_h defined by (3.3) is written as

$$a_h P_h / E(a_h) = P_h(1 + \varepsilon_h) \tag{A.1}$$

where

$$\varepsilon_h = \{a_h - E(a_h)\}/E(a_h), \tag{A.2}$$

we immediately observe, for any positive integer m, that

$$\varepsilon_h^m = (-1)^m (1 - a_h) + \left(\frac{Q_h^n}{1-Q_h^n}\right)^m a_h \tag{A.3}$$

where $Q_h = 1 - P_h$ and $E(a_h) = 1 - Q_h^n$ under the assumption that the strata sizes are sufficiently large for approximating the hypergeometric distribution of n_h by a binomial. Similarly the denominator of b_h can be written as

$$\sum_{k=1}^{L} a_k P_k / E(a_k) = \sum_{k=1}^{L} P_k(1 + \varepsilon_k) = 1 + \varepsilon \tag{A.4}$$

where $\varepsilon = \Sigma P_k \varepsilon_k$. It is not difficult to see that

$$E(\varepsilon_h) = 0, \; E(\varepsilon_h^m) = (-1)^m Q_h^n + Q_h^{nm} / (1 - Q_h^n)^{m-1} \sim O(Q_h^n),$$

$$E(\varepsilon) = 0, \; E(\varepsilon^m) = \sum_{r_1+\ldots+r_L=m} \binom{m}{r_1,\ldots,r_L} E\left(\prod_{k=1}^{L} P_k^{r_k} \varepsilon_k^{r_k}\right) \sim O(P_k^m Q_k^n). \tag{A.5}$$

Supposing that $|\varepsilon| < 1$, we are able to expand b_h in the form

$$b_h = P_h (1 + \varepsilon_h)(1 + \varepsilon)^{-1}$$

$$= P_h \{1 + (\varepsilon_h - \varepsilon) + (\varepsilon^2 - \varepsilon_h \varepsilon) + (\varepsilon_h \varepsilon^2 - \varepsilon^3) + \ldots\}, \tag{A.6}$$

from which we obtain

$$E(b_h - P_h) = P_h \{ E(\varepsilon^2 - \varepsilon_h \varepsilon) + E(\varepsilon_h \varepsilon^2 - \varepsilon^3) + \ldots\}. \tag{A.7}$$

Since it can be shown that $E(\varepsilon^m - \varepsilon_h \varepsilon^{m-1})$ is of order of magnitude $0(P_k^m Q_k^n)$ or $0(P_k^{m-1} Q_k^{n+1})$, the bias in $\tilde{R}$ is given by

$$E(\Sigma\, b_h \bar{y}_h) - \bar{Y} = \Sigma\{E(b_h) - P_h\}\, \bar{Y}_h \tag{A.8}$$

which is of order $0(P_h^2 Q_h^n)$ or $0(P_h Q_h^{n+1})$.

Using these results in (3.6) for computing the between strata component of variance we obtain

$$E(b_h)E(b_{h'}) - E(b_h b_{h'}) \sim 0(P_h^2 Q_h^n) \text{ or } 0(P_h Q_h^{n-1}). \tag{A.9}$$

We therefore conclude that $V(\tilde{R})_B \sim 0(P_h^2 Q_h^n)$ or $0(P_h Q_h^{n+1})$.

We thereby obtain a first approximation to $V(\tilde{R})_B$ to order $0(P_h^2 Q_h^n)$ or $0(P_h Q_h^{n+1})$ by omitting the terms in ε, ε_k, $\varepsilon_{k'}$ with degree higher than 2 in the expansion of

$$\begin{aligned} E(b_h)E(b_{h'}) - E(b_h b_{h'}) &= E[\{b_h - E(b_h)\}\{b_{h'} - E(b_{h'})\}] \\ &\doteq P_h P_{h'} E\{(\varepsilon_h - \varepsilon)(\varepsilon_{h'} - \varepsilon)\} \\ &= P_h P_{h'}\{V(\varepsilon) + \mathrm{Cov}(\varepsilon_h, \varepsilon_{h'}) - \mathrm{Cov}(\varepsilon_h, \varepsilon) - \mathrm{Cov}(\varepsilon_{h'}, \varepsilon)\} \end{aligned} \tag{A.10}$$

where

$$V(\varepsilon) = \sum_{k=1}^{L} \frac{1-E(a_k)}{E(a_k)} P_k^2 + \sum_{k' \neq k=1}^{L} \frac{E(a_k a_{k'}) - E(a_k)E(a_{k'})}{E(a_k)E(a_{k'})} P_k P_{k'}, \tag{A.11}$$

$$\mathrm{Cov}(\varepsilon_h, \varepsilon_{h'}) = \frac{E(a_h a_{h'}) - E(a_h)E(a_{h'})}{E(a_h)E(a_{h'})} \tag{A.12}$$

$$\mathrm{Cov}(\varepsilon_h, \varepsilon) = \frac{1-E(a_h)}{E(a_h)} P_h + \sum_{k \neq h} \frac{E(a_h a_k) - E(a_h)E(a_k)}{E(a_h)E(a_k)} P_k, \tag{A.13}$$

and

$$E(a_k) = 1 - \frac{\binom{N-N_k}{n}}{\binom{N}{n}}, \tag{A.14}$$

$$E(a_k a_{k'}) - E(a_k)E(a_{k'}) = \frac{\binom{N-N_k-N_{k'}}{n} - \binom{N-N_k}{n}\binom{N-N_{k'}}{n}}{\binom{N}{n}} . \qquad (A.15)$$

2. The order of magnitude of the bias, $V(\tilde{R})_B$ and $V(\tilde{R})_W$ when $L \to \infty$, $P_h \to 0$ and n is moderately large.

Since it is not difficult to prove that the bias and $V(R)_B$ are of the same order as before, we shall concentrate on $V(\tilde{R})_B$ and $V(\tilde{R})_W$. Without going into detail we can obtain a first approximation as follows:

$$V(\tilde{R}) = V(\tilde{y}/\tilde{x}) \doteq V(\tilde{y} - \bar{Y}\tilde{x}) \qquad (A.16)$$

$$= \sum_{h=1}^{L} \frac{V(a_h)}{\{E(a_h)\}^2} P_h^2 (\bar{Y}_h^2 - \bar{Y})^2$$

$$+ \sum_{h' \neq h=1}^{L} \frac{Cov(a_h, a_{h'})}{E(a_h)E(a_{h'})} P_h P_{h'} (\bar{Y}_h - \bar{Y})(\bar{Y}_{h'} - \bar{Y})$$

$$+ \sum_{h=1}^{L} \frac{\{E'(1/n_h)\, E(a_h) - E(a_h/N_h)\}}{\{E(a_h)\}^2} P_h^2 S_h^2 \qquad (A.17)$$

Assuming as before that the strata sizes are sufficiently large for approximation of the hypergeometric distribution by a multinomial, we can write

$$V(a_h) = Q_h^n (1 - Q_h^n), \qquad (A.18)$$

$$Cov(a_h, a_{h'}) = (1 - P_h - P_{h'})^n - Q_h^n Q_{h'}^n, \qquad (A.19)$$

and

$$E(a_h) = (1 - Q_h^n). \qquad (A.20)$$

We now consider cases where L is large, all P_h are small while n is moderate. Omitting all the terms in P_h with higher degree than 2 in the expansion of $V(a_h)$ and $Cov(a_h, a_h)$ and $E'(1/n_h)$ we obtain

$$V(a_h) \doteq n(P_h - P_h^2) - \frac{3}{2} n(n-1) P_h^2, \tag{A.21}$$

$$Cov(a_h, a_{h'}) \doteq -n P_h P_{h'} \tag{A.22}$$

and

$$E'(1/n_h) \doteq \frac{n P_h Q_n^{n-2}}{(1 - Q_h^n)} \left(Q_h + \frac{n-1}{2} P_h\right) \tag{A.23}$$

which on substituting in $V(\tilde{R})$ and omitting the finite population correction, reduce after much simplification to the form

$$V(\tilde{R}) = n \sum_{h=1}^{L} P_h \left[\frac{P_h \bar{Y}_h}{1 - Q_h^n} - \sum_{h=1}^{L} P_h \frac{P_h \bar{Y}_h}{(1-Q_h)^n}\right]^2$$

$$- \frac{3n(n-1)}{2} \sum_h \frac{P_h^4 \bar{Y}_h^2}{(1-Q_h^n)^2} + \sum_{h=1}^{L} \frac{nP_h^3 Q_h^{n-2} \left(Q_h + \frac{n-1}{2} P_h\right)}{(1 - Q_h^n)^2} \sigma_h^2. \tag{A.24}$$

We now infer that

$$V(\tilde{R})_B \sim O(n L^{-2}) \text{ or } O(n^2 L^{-3}), \tag{A.25}$$

and

$$V(\tilde{R})_W \doteq \sum_{h=1}^{L} \frac{P_h \sigma_h^2}{n} + O(L^{-1}) \tag{A.26}$$

which follows from

$$\frac{Q_h^{n-1}}{(1-Q_h^n)^2} = \frac{(1 - P_h)^{n-1}}{\{1-(1-P_h)^n\}^2} \doteq \frac{1-(n-1)P_h + \frac{(n-1)(n-2)}{2} P_h^2}{n^2 P_h^2 \{1-\frac{(n-1)}{2} P_h\}^2} \doteq \frac{1}{n^2 P_h^2}. \tag{A.27}$$

3. The order of magnitude of $V(\tilde{R})_W$ as $n \to \infty$

For large n it has been shown by Stephan (1945) that to terms of order n^{-2}

$$E'\left(\frac{1}{n_h}\right) \doteq \frac{1}{nP_h} + \frac{Q_h}{n^2 P_h}. \tag{A.28}$$

Then $$V(\tilde{R})_W \doteq \sum \frac{P_h \sigma_h^2}{n(1-Q_h^n)} + \frac{1}{n^2} \sum \frac{Q_h \sigma_h^2}{1-Q_h^n}. \tag{A.29}$$

APPENDIX II

An optimum property of $\tilde{R}$ for a population with equal strata sizes

We shall establish that when all strata sizes are equal $\tilde{R}$ is unbiased and minimizes uniformly the between strata component variance of a generalized class of unbiased ratio estimators

$$\tilde{R}_g = \frac{\Sigma\{g(n_h)P_h\bar{y}_h | E(g(n_h))\}}{\Sigma\{g(n_h)P_h | E(g(n_h))\}} \tag{A.30}$$

where $g(n_h)$ is any function with $g(0) = 0$. First of all these ratio estimators reduce to

$$\tilde{R}_g = \frac{\Sigma\, g(n_h)\bar{y}_h}{\Sigma\, g(n_h)} \tag{A.31}$$

since all P_h are equal. Moreover, the random variables

$$c_k = \frac{g(n_k)}{\Sigma\, g(n_h)}\ ,\ k=1,\ \ldots L, \tag{A.32}$$

have the same expectations, variances and covariances. Since

$$1 = \frac{\Sigma\, g(n_h)}{\Sigma\, g(n_h)} = \Sigma\ \frac{g(n_k)}{\Sigma\, g(n_h)} = \Sigma\, c_k \tag{A.33}$$

by taking the expectation and variance of this relation we arrive at

$$E(c_k) = 1/L\ ,\ \mathrm{Cov}(c_k, c_{k'}) = -V(c_k)/(L-1). \tag{A.34}$$

It follows that

$$E(\tilde{R}_g) = \underset{1}{E}\ (\underset{2}{E}(\Sigma c_h\bar{y}_h)) = \underset{1}{E}\ (\Sigma\, c_h\bar{y}_h) = \Sigma\ \bar{Y}_h/L\ = \bar{Y} \tag{A.35}$$

which implies that $\tilde{R}_g$ is unbiased. The between strata variance component can be treated in exactly the same way as $V(\tilde{R})_B$, we write

$$V(\tilde{R}_g)_B = -\sum_{h\neq h'} \frac{\mathrm{Cov}(c_h, c_{h'})}{2}\ (\bar{Y}_h - \bar{Y}_{h'})^2 = \frac{V(c_k)}{(L-1)}\ \Sigma\ \frac{(\bar{Y}_h - \bar{Y}_{h'})^2}{2} \tag{A.36}$$

which, after some simplification,

$$= V(c_k)\ L\ S_B^2 \tag{A.37}$$

where

$$S_B^2 = \frac{\Sigma \bar{Y}_h^2 - L\ \bar{Y}^2}{(L-1)} \tag{A.38}$$

Since

$$V(c_k) = E(c_k^2) - \{E(c_k)\}^2 = E(c_k^2) - 1/L^2, \tag{A.39}$$

we concentrate on minimizing

$$E(c_k^2) = \frac{1}{L}\ \Sigma\ E(c_h^2) = E\left(\frac{\Sigma c_h^2}{L}\right) = \frac{1}{L}\ \Sigma\ \left(\frac{\Sigma\{g(n_h)\}^2}{\{\Sigma g(n_h)\}^2}\right) P(n_1, \ldots, n_L) \tag{A.40}$$

where the first summation extends over all possible values of $n_1, \ldots, n_L$ and $P(n_1, \ldots, n_L)$ is the probability of getting $(n_1, \ldots, n_L)$ in a sample of size n. For any particular value $(n_1', \ldots, n'_L)$ with ν positive n_h values, we can see, by Cauchy-Schwartz inequality,

$$\frac{\Sigma\{g(n_h')\}^2}{\{\Sigma g(n_h')\}^2} \geq \frac{1}{\nu} \tag{A.41}$$

and equality is attained when $g(n_h')$ = arbitrary nonzero constant. Without loss of generality we assume $g(n_h) = 1$ for all $n_h \neq 0$ which minimize $E(c_k^2)$, i.e. $\tilde{R}$ minimizes the between stratum component of variance. In fact

$$V(\tilde{R})_B = \{E(\frac{1}{\nu}) - \frac{1}{L}\}S_B^2 \tag{A.42}$$

References

Hartley, H. O. (1959). "Analytic studies of survey data." Special publication by The Inst. of Statist. of U. of Rome in honor of Prof. C. Gini, 1-32.

Hartley, H. O., Rao, J. N. K., and Kiefer, J. (1969). "Variance estimation with one unit per stratum." J. Am. Statist. Assoc. 64, 841-851.

Horvitz, D. G., and Thompson, D. J., (1952). "A generalization of sampling without replacement from a finite universe." J. Am. Statist. Assoc., 47, 663-685.

Stephan, F. F. (1945). "The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate." Ann. Math. Statist., 16, 50-61.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

BEST AVAILABLE COPY

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (19) 14209.1-M / (18) ARO	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) (6) SAMPLE An Exact Small Theory for Post-Stratification	5. TYPE OF REPORT & PERIOD COVERED (9) Technical Report
6. PERFORMING ORG. REPORT NUMBER (14) TR-1	
7. AUTHOR(s) (10) D. C. Doss, H. O. Hartley, G. R. Somayajulu	8. CONTRACT OR GRANT NUMBER(s) (15) DAAG29-77-G-0086
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University College Station, Texas 77843	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709	12. REPORT DATE (11) March 1978
	13. NUMBER OF PAGES 21 (12) 24p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A genuine small sample theory for post-stratification is developed in this paper. This includes the definition of a ratio estimator of the population mean $\bar{Y}$, the derivation of its bias and its exact variance and a discussion of variance estimation. The estimator has both a within strata component of variance which is comparable with that obtained in proportional allocation stratified sampling and a between strata component of variance which will tend to zero as the overall sample size becomes large. Certain optimality properties of the estimator are obtained. The generalization of post-stratification from the simple random sampling to post-stratification used in →

next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

347 350

Unclassified

20. ABSTRACT CONTINUED

conjunction with stratification and multi-stage designs is discussed.